结合状态预测的深度强化学习交通信号控制 *

唐慕尧，周大可†，李　涛

(南京航空航天大学 自动化学院，南京 211100)

摘　要： 深度强化学习(Deep Reinforcement Learning，DRL)可广泛应用于城市交通信号控制领域，但在现有研究中，绝大多数的 DRL 智能体仅使用当前的交通状态进行决策，在交通流变化较大的情况下控制效果有限。文中提出一种结合状态预测的 DRL 信号控制算法。首先，利用独热编码设计简洁且高效的交通状态；然后，使用长短期记忆网络(Long Short-Term Memory，LSTM)预测未来的交通状态；最后，智能体根据当前状态和预测状态进行最优决策。在 SUMO(Simulation of Urban Mobility)仿真平台上的实验结果表明，在单交叉口、多交叉口的多种交通流量条件下，与三种典型的信号控制算法相比，所提算法在平均等待时间、行驶时间、燃油消耗、CO2 排放等指标上都具有最好的性能。

关键词： 交通信号控制；状态预测；深度强化学习；深度 Q 网络；长短期记忆网络

中图分类号： TP181　　doi: 10.19734/j.issn.1001-3695.2021.12.0704

State prediction based deep reinforcement learning for traffic signal control

Tang Muyao, Zhou Dake†, Li Tao

(*School of Automation Engineering*, *Nanjing University of Aeronautics & Astronautics*, *Nanjing* 211100, *China*)

Abstract: Urban traffic signal control can widely use deep reinforcement learning (DRL) technique. However, in existing researches, most DRL agents only use the current traffic state to make decisions and have limited control effects when the traffic flow changes greatly. Aiming at the problem, this paper proposed a state prediction based deep reinforcement learning algorithm for traffic signal control. The algorithm used one-hot coding to design a concise and efficient traffic state, and then used a Long Short-Term Memory (LSTM) to predict the future state. The agent made optimal decisions based on the current state and the predicted state. The experimental results on the simulation platform SUMO show that compared with three typical signal control algorithms, the proposed algorithm has the best performance in terms of average waiting time, travel time, fuel consumption, CO2 emissions and cumulative reward both in a single intersection and multiple intersections under different flow conditions.

Key words: traffic signal control; state prediction; deep reinforcement learning; deep q network; long short-term memory

0　引言

随着人们生活水平的提高，汽车保有量持续增长，城市的交通拥堵问题也日趋严重。交通信号控制是提高道路通行效率、缓解交通拥堵最直接、成本最低的途径。SCATS[1]和 SCOOT[2]是目前广泛使用的自适应交通信号控制系统，前者选择信号配时方案，后者利用简化的交通模型求解最优的信控策略。但是，简化模型的建立依赖于大量的假设和经验方程，因此，对于复杂多变的真实交通场景，这类系统的效果欠佳。近年来，随着人工智能技术的发展，强化学习[3](Reinforcement Learning，RL)尤其是数据驱动的深度强化学习，在交通信号控制方面展现出卓越的应用前景。

强化学习是一种“试错”的学习方法，通过与环境交互来学习最优策略。应用在交通信号控制中，可以把一个或几个交叉口看成一个智能体(Agent)，智能体观测路网状态后作出决策，通过最大化环境反馈的奖励以学习最优的信号配时方案。受到人脑工作模式的启发，深度学习[4](Deep Learning，DL)能够把底层特征组合形成更加抽象的高层特征，可以有效处理高维数据。深度强化学习(DRL)结合了 DL 的强感知能力与 RL 的强决策能力，非常适用于交通信号控制的任务。

2010 年，Arel 等[5]首次将 DRL 引入交通信号控制领域，使用神经网络拟合 Q 值，但是缺少经验回放、目标网络部分。Liu 等[6]提出 3DQN_PSER 算法，使用优先级序列经验回放(Priority Sequence Experience Replay，PSER)更新经验池中序列样本的优先级，使智能体获取与交通状态相似的前序样本，提高训练效率。Wei 等[7]提出模型 Intellilight，使用相位门结构设置独立的学习通道，根据相位、动作对经验池进行划分，并用真实的交通数据做实验。Zheng 等[8]提出 FRAP 模型，利用不同信号相位间的竞争关系，实现了在交通流中翻转和旋转等对称情况下的普适性。Jin 等[9]使用动作策略阈值词典排序法(Threshold Lexicographic Ordering，TLO)自适应地选择优化目标，基于 SARSA 算法对比多种函数逼近方法的改善效果。Tan 等[10]将大规模路网分为若干个子区域，对每个区域，使用 Peraction DQN 或 Wolpertinger DDPG 进行控制，将所有智能体的学习策略传递给全局智能体实现全局学习。这些 DRL 信控方法本质上是一阶马尔可夫决策过程，智能体仅根据当前的状态进行决策，在复杂多变的实际交通场景下难以实现最优的控制效果。如果能合理预测未来状态，智能体将提前考虑可能出现的交通情况，学习更好的信控策略。Xu 等[11]提出了 DRQN 模型，跨 8 个时间步长集成隐藏状态

收稿日期： 2021-12-26；**修回日期：** 2022-03-21　**基金项目：** 国家自然科学基金资助项目(62073164)；南京航空航天大学研究生创新基地(实验室)开放基金资助项目(kfjj20200313)

作者简介： 唐慕尧(1997-)，男，江苏泰州人，硕士研究生，主要研究方向为智能控制；周大可(1974-)，男(通信作者)，江苏淮安人，副教授，硕导，博士，主要研究方向为机器学习、计算机视觉与智能控制等(dkzhou@nuaa.edu.cn)；李涛(1979-)，男，安徽淮南人，副教授，硕导，博士，主要研究方向为网络化多智能体系统、网络控制系统与飞行器控制.

输入 DRL 智能体，但这样显著地增加了状态的维数，容易导致神经网络过拟合。循环神经网络具有短时记忆能力，Chu 等[12]在 DRL 智能体中采用 LSTM 网络来提取动态的交通信息，但该网络并没有直接预测未来的交通状态。

本文提出了一种结合状态预测的深度强化学习信号控制算法 DQN_SP，主要特点有：1)通过引入显式的交通状态预测，DRL 智能体利用当前和未来状态进行最优决策。2)精心设计智能体的状态，该状态包含最重要的交通信息且数据量小易于预测。在单交叉口、多交叉口的多种流量条件下验证了所提算法的有效性与可行性，车流数据模拟了现实中高低峰的情形，具有工程应用价值。

1 研究背景

本节将介绍强化学习、深度强化学习的基本概念和方法，以及 DRL 信号控制算法。

1.1 强化学习

强化学习是和有监督学习、无监督学习并列的第三类机器学习方法，智能体通过与环境不断交互来学习为了达成某个目标所需的最佳策略。马尔可夫决策过程是一种通过交互式学习来达到目标的理论框架，其灵活抽象，可以很好地解释强化学习的基本流程。智能体根据当前策略，以一定概率执行最优动作并与环境交互，用动作价值函数 $q_\pi(s,a)$ 来表示智能体在状态 s 下采取动作 a 的期望回报，表示为

$$q_\pi(s,a) \doteq E_\pi[G_t \mid S_t = s, A_t = a] = E_\pi[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \mid S_t = s, A_t = a] \tag{1}$$

智能体在与环境交互后学习到最优策略，最优动作价值函数为在状态 s 下采取动作 a 获得的最高回报值，根据贝尔曼最优方程，可得：

$$q_*(s,a) = E[R_{t+1} + \gamma \max_{a'} q_*(s_{t+1},a') \mid S_t = s, A_t = a] = \sum_{s',r} p(s',r \mid s,a)[r + \gamma \max_{a'} q_*(s',a')] \tag{2}$$

不断迭代最优动作价值函数 $q_*(s,a)$ 后，得到最优策略：

$$\pi_* = \arg\max_{a \in A} q_*(s,a) \tag{3}$$

1.2 深度强化学习

DRL 是 RL 与 DL 的结合，是目前控制系统中先进的学习框架之一。2013 年 DeepMind[13]提出了 DQN，不同于 Q-Learning 使用一张表来保存所有的 Q 值，DQN 使用经验回放来更新目标价值。将智能体与环境交互获得的样本 (s,a,r,s') 存入经验池中，从经验池均匀采样小批量样本，使用随机梯度下降方法训练深度神经网络使其逼近 Q 值，随机采样能够打破样本间的强相关性，使训练收敛稳定。DQN 使用同一个网络来选择动作和计算目标 Q 值，两者在迭代的过程中相互依赖，不利于算法的收敛。为了解决此问题，DeepMind 提出了 Nature DQN[14]，使用两个网络，当前网络 Q 用来选择动作、更新参数，目标网络 Q^- 用来计算目标 Q 值，Q^- 网络的参数不需要实时迭代更新，而是每隔一段时间从当前网络 Q 复制过来，当前最优动作价值函数的优化目标表示为

$$y(s,a) = r + \gamma \max_{a'} q(s',a';w^-) \tag{4}$$

其中 w^- 表示目标值网络的参数。

上述算法计算目标 Q 值时都是通过贪婪法直接得到，虽然可以快速让 Q 值向优化目标靠近，但是很容易导致过度估计。为了缓解模型的过拟合问题，Van Hasselt 等人[15]提出了 Double DQN，先在当前网络 Q 中寻找最大 Q 值对应的动作，再将此动作代入目标网络 Q^- 计算目标 Q 值，优化目标表示为

$$y(s,a) = r + \gamma q(s', \arg\max_{a'} q(s',a';w); w^-) \tag{5}$$

上述算法通过经验回放来训练深度 Q 网络，在经验池中对样本均匀采样，然而不同样本 TD 误差不同，对反向传播的影响也不一样。为了解决此问题，Schaul 等[16]基于 DDQN 提出了优先经验回放算法，给定正比于样本 TD 误差绝对值 $|\delta(t)|$ 的优先级，并将优先级存入经验池，训练时使优先级高的样本更容易被采样，避免没有价值的迭代，提高算法收敛速度。Wang 等[17]对神经网络结构进行优化，提出 Dueling DQN，将 Q 网络分为价值函数与优势函数两部分。

1.3 基于 DRL 的交通信号控制

基于 DRL 的信控方法不需要场景先验知识，而是通过与交通环境不断交互来学习最优策略。在此过程中，交叉口或路网看成一个智能体，状态是对交通环境的描述，动作是交通信号的变化，奖励衡量了动作后交通通行的效率变化。

现有的 DRL 信控算法在状态、动作、奖励定义方面有很大不同。状态定义分为两类，基于车辆的表示(如实时图像[7, 18]、包括车辆位置或速度信息的 DTSE 形式[6, 19, 20])，和基于特征的值向量表示(如排队长度[7, 19, 21]、累计延误[19, 20]、等待时间[7, 19])。动作定义分为选择一个可能的绿灯相位[6, 20, 21]、保持当前相位或切换至下一相位[7, 11, 19]，或改变相位持续时间[9, 22]。状态是环境的特征矩阵或向量，动作是离散的选择向量，奖励是与交通数据有关的标量值。奖励的定义主要考虑队列长度[6, 7, 19, 20]、延误[7, 19, 20, 22]等。DRL 算法主要分为基于值函数的 DQN[6, 7, 11, 19, 20]、基于策略的 DDPG[10, 23]、基于 AC 框架的 A2C[12, 18]、A3C[24]等。

一些研究考虑了交通流的时序相关性。Yu 等[23]把车辆的速度加入到状态表示中，Wei 等[7]把表示车辆位置的实时图像喂入 CNN 网络，这两种方法通过合理设计状态体现交通流的动态特性。Chu 等[12]使用 LSTM 网络拟合 Q 值，利用网络的记忆能力学习交通信息的变化趋势，但是没有直接预测未来的交通状态。为了克服 DQN 无法记住当前输入之前的历史信息这一缺点，Xu 等[11]提出了 DRQN 模型，将当前的状态和几个历史状态输入智能体，可以看成 n 阶马尔可夫决策过程。Liu 等[6]使用 PSER 更新经验池中序列样本的优先级，使当前时刻之前的样本数据更容易被采样。上述方法或多或少考虑了交通流的时序特性，但是没有对交通状态直接预测，因为微观状态维数大，容易引发维数灾难的问题，且和 DRL 结合时难以训练出令人满意的结果。

2 结合状态预测的深度强化学习交通信号控制算法

本文将状态预测与 DRL 中的 DQN 算法相结合，采用独热编码的形式精心设计微观状态，并用 LSTM 预测未来的状态，智能体根据当前状态和预测状态进行决策。本节将对状态、动作、奖励进行定义，并介绍所提算法 DQN_SP 的网络模型。

2.1 状态定义

本文需要利用当前和预测的交通状态进行决策，因此状态设计尤为关键。基于 DTSE 方法采用非均匀量化和独热编码来设计状态向量。本文用于仿真的交叉口为双向 6 车道，长 500 米，沿着车辆的行驶方向，左边的车道为左转车道，中间车道为直行车道，右边的车道为直行加右转车道。本文按照一定长度比例将车道划分为元胞，图 1 所示的是以交叉口西进口道为例的元胞设计图。其中，右边的两条车道看做一个整体进行划分，左边的左转车道单独进行划分，这样一个交叉口四个方向的车道将被划分为 80 个元胞。状态由每个元胞中是否有车辆来表示，如有车辆，状态取值为 1，否则为 0。

由图 1 西进口道的元胞设计图可以看出，交叉口附近以 7 米为单位划分出 10 个元胞，其中每个元胞都只能容纳 1 辆车，可以精确地反映车辆分布情况，离交叉口最远的元胞长 230 米。与用实时图像[18]或对车道均匀划分[19]表示状态的方

法相比，该方法使智能体更关注靠近路口的交通状况，降低了数据维度，缩短了计算时间。以每个元胞中是否有车辆作为状态，简化交通信息，能够反映环境的主要特征，即交叉口附近的车辆分布情况。另外，对这种独热编码形式的状态进行预测，可以看成二分类问题，相比于传统的回归预测，能够提高预测准确率。

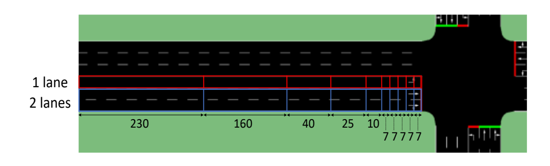


图 1　交叉口西进口道的元胞设计图

Fig. 1　Cellular design of the west entrance road at the intersection

2.2　动作定义

智能体需要根据交通状态选择合适的动作来疏导交通，本文的动作定义为选择一个可能的信号相位。动作集合 $A=\{NSG, NSLG, EWG, EWLG\}$，分别表示南北方向直行和右转绿灯、南北方向左转绿灯、东西方向直行和右转绿灯、东西方向左转绿灯。每个相位的最短持续时间设为 10 s，同时为了安全起见，绿灯和红灯切换期间会有时长 3 s 的黄灯。

2.3　奖励定义

智能体在 t 时刻观测环境状态为 s_t，执行动作 a_t 后得到环境对该动作的反馈 r_t，用来衡量该动作的质量，是 DRL 能否收敛以及能否取得良好效果的关键。本文奖励 r_t 定义为相邻时间步的所有车道车辆排队长度之差：

$$r_t = \alpha q_t - q_{t+1} \tag{6}$$

其中 q_t 表示 t 时刻路网中所有车道的排队长度之和，q_{t+1} 表示下一时间步所有车道的排队长度之和，α 为系数，通过多次实验后设为 0.9。

2.4　结合状态预测的 DRL 信控算法(DQN_SP)

本文所提算法 DQN_SP 采用 LSTM 预测未来微观状态 s_p，并将其与当前状态 s 串联，作为增广状态输入 DRL 智能体，DRL 算法使用传统的 DQN[13]，旨在验证结合状态预测后算法的有效性与可行性。DQN_SP 的网络结构如图 2 所示，最优动作价值函数的优化目标表示为

$$y(s,s_p,a) = r + \gamma \max_{a'} q(s', s_p', a') \tag{7}$$

图 2　DQN_SP 的网络结构

Fig. 2　The network structure of DQN_SP

DQN_SP 的算法流程如下所示。

```
初始化深度 Q 网络、LSTM 网络、经验池
for episode = 1 to M do
  初始化路网环境，导入车流数据
  for t = 1 to T do
    智能体观测当前环境状态 s
    LSTM 预测 n 个时间步后的微观状态 s_p
    当前状态与预测状态串联后输入 DQN 智能体，智能体基于 ε 贪婪策略执行动作 a
    智能体进入新的状态 s'，根据式(6)计算奖励 r
    LSTM 预测 n 个时间步后的微观状态 s_p'
    将样本 (s,s_p,a,r,s',s_p') 存入经验池中
  end for
  从经验池中抽取样本训练网络
  根据式(7)计算优化目标，使用均方误差损失函数更新深度 Q 网络参数 w
  使用二值交叉熵损失函数更新 LSTM 网络参数 θ
end for
```

3　实验结果与分析

本节首先介绍实验的仿真环境与算法超参数，然后介绍基准算法 FTC、SOTL、DQN，最后在单交叉口、多交叉口的多种流量条件下验证了算法 DQN_SP 的有效性。

3.1　仿真环境与超参数设置

SUMO 是免费开源的交通系统仿真软件，其中的 Traci(Traffic Control Interface)接口可以与多种开发环境在线交互，实现对交通信号的控制。本文以 Ubuntu GeForce RTX 2080 GPU 作为硬件环境，算法通过深度学习框架 Keras 实现，在 SUMO v1.6.0 下进行仿真实验。

交叉口设置：本文在单交叉口和多交叉口两种场景下分别进行仿真。交叉口由 4 条垂直的道路组成，每条道路长 500 米，为双向六车道，沿着车辆的行驶方向左边为左转车道，中间为直行车道，右边为直行加右转车道。多交叉口为 4 个相同的交叉口组成的 2×2 井字形路网，路口配置同单交叉口。

交通流设置：车辆生成的方式对交通信号控制有着重要的影响，本文中车辆的生成服从韦伯分布，其概率密度函数为

$$f(x;\lambda,a)=\begin{cases}\dfrac{a}{\lambda}\left(\dfrac{x}{\lambda}\right)^{a-1}e^{-\left(\frac{x}{\lambda}\right)^{a}} & x\ge 0\\ 0 & x<0\end{cases} \tag{8}$$

其中 λ 是比例参数设为 1，a 是形状参数设为 2，绝大多数车辆集中在某一段时间内进入路网，可以模拟现实生活中高峰低峰的情形。车辆从任意入口进入路网，以 75%的概率直行，12.5%的概率左转，12.5%的概率右转。车辆长 5 m，加速度为 1 m/s^2，以 36 km/h 的速度进入路网，最大速度为 50 km/h，车辆之间最小间距为 2.5 m。

超参数设置：参照文献[7, 9, 19]并结合实验，超参数设置如下。训练回合数设为 100，算法使用 DNN 评估 Q 值，隐藏层数为 5，宽度为 400，采用 Adam 优化器，学习率为 0.001，批处理大小为 80，每回合训练迭代 800 次，采用均方误差作为损失函数。预测网络使用 6 个 LSTM 单元，每个单元有 3 个 LSTM 层，神经元个数为 80，采用 Adam 优化器，批处理大小为 128，每回合训练迭代 1 次，采用二值交叉熵作为损失函数。RL 经验池尺寸最小为 600，最大为 50000，折扣因子为 0.75，使用 ε 贪婪算法输出动作。

3.2　实验评估与结果分析

本文在单交叉口和多交叉口两种场景下分别实验。对于单交叉口，仿真时长为 5400 s，进入路网的车辆数目为 500、1000、1500，分别对应低、中、高三种流量条件。对于多交叉口，仿真时长也为 5400 s，进入路网的车辆数目设为 2000、3000，分别对应低、高两种流量条件。对于每种流量条件，用随机种子 seed 生成 20 组车流数据，20 组数据下车辆的平均等待时间、平均行驶时间、平均燃油消耗、平均 CO_2 排放、平均累计奖励作为算法的性能指标。其中，平均等待时间主

要来自于车辆排队时消耗的时间，与定义的奖励相关性最强，为主要指标，平均行驶时间、燃油消耗、CO_2 排放为次要指标。所提算法对 1、5、10 个时间步后的状态进行预测，分别记为 DQN_SP_1、DQN_SP_5、DQN_SP_10。为了验证预测的有效性，将 DQN_SP 与下列基准算法进行比较：

固定配时控制(Fixed-time Control，FTC)。FTC 根据经典的韦伯斯特配时法[25]预先定义一套配时方案，广泛应用于现实交通场景中。

自组织交通灯(Self-organizing Traffic Lights，SOTL)[26]。当红灯方向的排队长度达到阈值时，该方向的信号灯就变成绿灯，若绿灯方向一定距离内车辆数过多，则延长绿灯时长。

基于 DQN 的交通信号控制。使用与所提算法 DQN_SP 相同的 DQN 算法[13]，唯一区别在于其不对未来状态进行预测，所以网络输入维度减半，其余超参数设置以及状态、动作、奖励定义与 DQN_SP 相同。

图 3 是在单交叉口中流量条件下，训练与测试过程中，各算法的累计奖励对比和车辆平均等待时间对比。图 3(a)给出了在单交叉口中流量条件下，DQN_SP 与 DQN 在训练过程中的累计奖励对比，两者区别不大。可见，增加了状态预测，不会降低算法的收敛速度，也不会削弱算法稳定性。图 3(b)表示 DQN_SP 与三种基准算法的车辆平均等待时间对比。在训练的初始阶段，由于经验池中的样本太少，智能体还没有学到正确的控制策略，所以平均等待时间会大幅上升，随着训练的进行，交叉口通行状况逐渐好转，最终趋于平稳。

训练好的模型在随机生成的 20 组车流数据下进行测试，平均性能如表 1 所列，可以看出无论是预测 1 步、5 步还是 10 步后的状态，DQN_SP 的性能都比 FTC、SOTL、DQN 更加优越，且在主要指标上，DQN_SP_5 改善最多，相比于 DQN，车辆平均等待时间减少了 6.06%，累计奖励提高了 5.61%。然而在行驶时间、燃油消耗、CO_2 排放这三个次要指标上，DQN_SP_1 改善效果最明显。图 3(c)表示 DQN_SP_5 与 DQN 在 20 次测试中的累计奖励对比，图 3(d)表示 DQN_SP_5 与三种基准算法的车辆平均等待时间对比。结果显示，相较于传统的 FTC、SOTL 信号控制，基于 DRL 的方法在缩短车辆的等待时间上效果显著，且在 18 次测试中，DQN_SP_5 的控制效果均优于 DQN。

表 2 单交叉口低流量条件下算法的性能

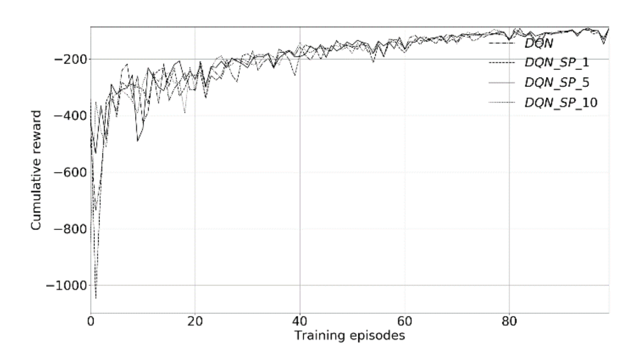
Tab. 2 Performance of algorithms under the condition of low traffic flow at an intersection

Algorithm	Waiting time/s	Travel time/s	Fuel consumption/ml	CO_2 emissions/g	Cumulative reward
FTC	17.73	101.64	87.99	204.69	\
SOTL	8.78	92.31	77.86	181.12	\
DQN	7.98	90.82	76.49	177.93	-37.11
DQN_SP_1	7.57	90.29	75.78	176.29	-35.54
DQN_SP_5	**7.41**	**89.97**	**75.41**	**175.42**	**-34.59**
DQN_SP_10	7.73	90.45	75.94	176.67	-36.07

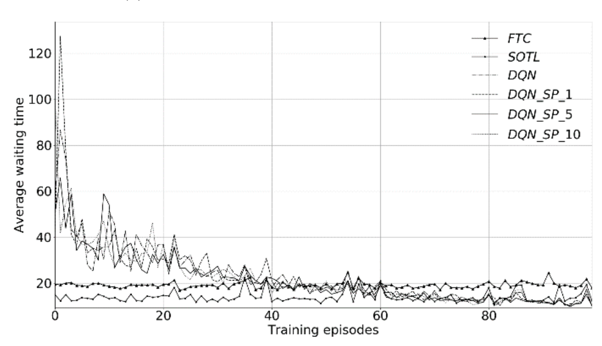
表 3 单交叉口高流量条件下算法的性能

Tab. 3 Performance of algorithms under the condition of high traffic flow at an intersection

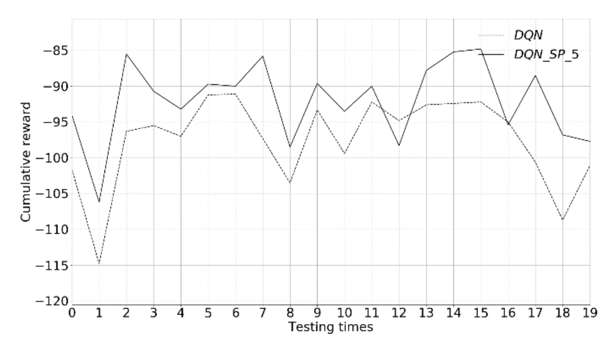
Algorithm	Waiting time/s	Travel time/s	Fuel consumption/ml	CO_2 emissions/g	Cumulative reward
FTC	22.75	109.20	95.60	222.40	\
SOTL	25.13	114.17	99.06	230.45	\
DQN	16.16	102.67	88.72	206.39	-207.04
DQN_SP_1	15.40	101.88	87.85	204.37	-197.53
DQN_SP_5	15.15	101.53	87.58	203.73	-194.86
DQN_SP_10	**14.68**	**101.09**	**87.08**	**202.58**	**-189.16**



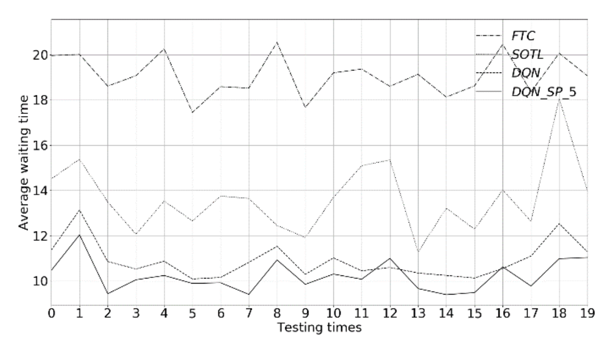
(a) 训练过程中各算法的累计奖励对比



(b)训练过程中各算法的车辆平均等待时间对比



(c)测试过程中算法 DQN 和 DQN_SP_5 的累计奖励对比



(d)测试过程中各算法的车辆平均等待时间对比

图 3 各算法的累计奖励对比和车辆平均等待时间对比

Fig.3 Comparison of cumulative rewards of algorithms and comparison of average waiting time of vehicles

本文还在多交叉口场景下进行实验，每个交叉口信号都用一个智能体控制。本文旨在验证结合状态预测的 DRL 的有效性，因此，使用简单的多智能体协作策略：采用空间折扣因子削弱来自其他交叉口的奖励，当前交叉口奖励权重为 0.5，邻居交叉口为 0.2，对角交叉口为 0.1。仿真时长 5400 秒，进入路网的车辆数目设为 2000、3000 辆，分别对应低流量和高流量，表 4、5 列出了算法在 20 次测试中的平均性能。在高流量情况下，SOTL 控制效果糟糕，因为当交通流高度随机的时候，车辆驱动的控制方法很难奏效。在低流量条件下，DQN_SP_5 的改善效果最好，相比于 DQN，平均等待时间减少 8.82%，累计奖励提高 8.11%，然而在高流量条件下，

DQN_SP_10 的改善效果最好，平均等待时间减少 4.92%，累计奖励提高 4.59%。由此可见，随着车流量变大，需要对更多时间步后的状态进行预测，以更有效地学习交通变化趋势，提高通行能力。

表 4　多交叉口低流量条件下算法的性能

Tab. 4　Performance of algorithms under the condition of low traffic flow at multiple intersections

Algorithm	Waiting time/s	Travel time/s	Fuel consumption/ml	CO_2 emissions/g	Cumulative reward
FTC	45.05	177.37	157.77	367.03	\
SOTL	29.15	158.72	136.97	318.63	\
DQN	21.54	151.11	129.86	302.11	-371.72
DQN_SP_1	20.23	149.39	128.01	297.79	-350.65
DQN_SP_5	**19.64**	**148.78**	**127.39**	**296.34**	**-341.59**
DQN_SP_10	20.27	149.46	128.06	297.92	-351.12

表 5　多交叉口高流量条件下算法的性能

Tab. 5　Performance of algorithms under the condition of high traffic flow at multiple intersections

Algorithm	Waiting time/s	Travel time/s	Fuel consumption/ml	CO_2 emissions/g	Cumulative reward
FTC	51.05	185.78	166.42	387.14	\
SOTL	61.47	204.15	178.84	416.02	\
DQN	34.36	167.89	147.45	343.02	-836.09
DQN_SP_1	33.23	166.81	146.37	340.51	-811.05
DQN_SP_5	32.87	166.60	146.19	340.08	-803.14
DQN_SP_10	**32.67**	**166.59**	**146.02**	**339.68**	**-797.98**

综上所述，相较于基准算法，DQN_SP 在单交叉口和多交叉口的场景下都能学习更好的信号控制策略，有效缓解了交通拥堵，减少燃油消耗与污染排放。随着车流量的增多，需要预测更多时间步后的状态以获得更好的控制效果。

4　结束语

本文利用了交通数据的时序相关性，提出结合状态预测的深度强化学习交通信号控制算法 DQN_SP，通过提取高维交通特征，并对未来微观状态进行预测，在单交叉口、多交叉口以及多种流量条件下都取得了更好的信控效果。与 FTC、SOTL、DQN 算法相比，DQN_SP 在平均等待时间、行驶时间、燃油消耗、CO_2 排放方面具有提升。未来本文将进一步研究将状态预测与更先进的 DRL 算法(如 TD3、SAC 等)相结合，并使用真实的交通数据进行验证。

参考文献：

[1] Sims A G, Finlay A B. SCATS, splits and offsets simplified (SOS) [J]. Australian Road Research, 1984, 12 (4): 17-33.

[2] Hunt P B, Robertson D I, Bretherton R D, *et al.* The SCOOT on-line traffic signal optimisation technique [J]. Traffic Engineering & Control, 1982, 23 (4): 190-192.

[3] Sutton R S, Barto A G. Reinforcement learning: an introduction [M]. MIT Press, 2018.

[4] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521 (7553): 436-444.

[5] Arel I, Liu C, Urbanik T, *et al.* Reinforcement learning-based multi-agent system for network traffic signal control [J]. IET Intelligent Transport Systems, 2010, 4 (2): 128-135.

[6] 刘志, 曹诗鹏, 沈阳, 等. 基于改进深度强化学习方法的单交叉口信号控制 [J]. 计算机科学, 2020, 47 (12): 226-232. (Liu Zhi, Cao Shipeng, Shen Yang, *et al.* Signal control of single intersection based on improved deep reinforcement learning method [J]. Computer Science, 2020, 47 (12): 226-232.)

[7] Wei Hua, Zheng Guanjie, Yao Huaxiu, *et al.* Intellilight: a reinforcement learning approach for intelligent traffic light control [C]// Proc of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2018: 2496-2505.

[8] Zheng Guanjie, Xiong Yuanhao, Zang Xinshi, *et al.* Learning phase competition for traffic signal control [C]// Proc of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2019: 1963-1972.

[9] Jin Junchen, Ma Xiaoliang. A multi-objective agent-based control approach with application in intelligent traffic signal system [J]. IEEE Trans on Intelligent Transportation Systems, 2019, 20 (10): 3900-3912.

[10] Tan Tian, Bao Feng, Deng Yue, *et al.* Cooperative deep reinforcement learning for large-scale traffic grid signal control [J]. IEEE Trans on Cybernetics, 2019, 50 (6): 2687-2700.

[11] Xu Ming, Wu Jianping, Huang Ling, *et al.* Network-wide traffic signal control based on the discovery of critical nodes and deep reinforcement learning [J]. Journal of Intelligent Transportation Systems, 2020, 24 (1): 1-10.

[12] Chu Tianshu, Wang Jie, Codecà L, *et al.* Multi-agent deep reinforcement learning for large-scale traffic signal control [J]. IEEE Trans on Intelligent Transportation Systems, 2019, 21 (3): 1086-1095.

[13] Mnih V, Kavukcuoglu K, Silver D, *et al.* Playing atari with deep reinforcement learning [J]. arXiv preprint arXiv: 1312. 5602, 2013.

[14] Mnih V, Kavukcuoglu K, Silver D, *et al.* Human-level control through deep reinforcement learning [J]. Nature, 2015, 518 (7540): 529-533.

[15] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning [C]// Proc of the 30th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2016, 30 (1): 2094-2100.

[16] Schaul T, Quan J, Antonoglou I, *et al.* Prioritized experience replay [C]// Proc of the 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016: 322-355.

[17] Wang Z, Schaul T, Hessel M, *et al.* Dueling network architectures for deep reinforcement learning [C]// Proc of the 33rd International Conference on Machine Learning. New York: ACM Press, 2016: 1995-2003.

[18] Mousavi S S, Schukat M, Howley E. Traffic light control using deep policy-gradient and value-function-based reinforcement learning [J]. IET Intelligent Transport Systems, 2017, 11 (7): 417-423.

[19] 孙浩, 陈春林, 刘琼, 等. 基于深度强化学习的交通信号控制方法 [J]. 计算机科学, 2020, 47 (2): 169-174. (Sun Hao, Chen Chunlin, Liu Qiong, *et al.* Traffic signal control method based on deep reinforcement learning [J]. Computer Science, 2020, 47 (2): 169-174.)

[20] Van der Pol E, Oliehoek F A. Coordinated deep reinforcement learners for traffic light control [C]// Proc of the 30th Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2016: 1-9.

[21] Wang Xiaoqiang, Ke Liangjun, Qiao Zhimin, *et al.* Large-scale traffic signal control using a novel multiagent reinforcement learning [J]. IEEE Trans on Cybernetics, 2020, 51 (1): 174-187.

[22] Touhbi S, Babram M A, Nguyen-Huu T, *et al.* Adaptive traffic signal control: exploring reward definition for reinforcement learning [J]. Procedia Computer Science, 2017, 109: 513-520.

[23] Yu Bingquan, Guo Jinqiu, Zhao Qinpei, *et al.* Smarter and safer traffic signal controlling via deep reinforcement learning [C]// Proc of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2020: 3345-3348.

[24] Genders W, Razavi S. Evaluating reinforcement learning state

representations for adaptive traffic signal control [J]. Procedia Computer Science, 2018, 130: 26-33.

[25] Webster F V. Traffic signal settings, road research technical [J]. Road Research Laboratory, 1958, 39.

[26] Cools S B, Gershenson C, D'Hooghe B. Self-organizing traffic lights: a realistic simulation [M]// Advances in Applied Self-organizing Systems. London: Springer, 2013: 45-55.